Statistical Tests for the Detection of Single Nucleotide Polymorphisms Using DNA Pooling

Ramsey, David M.
University of Limerick, Department of Mathematics and Statistics
Plassey
Limerick, Ireland
E-mail: david.ramsey@ul.ie

Futschik, Andreas
University of Vienna, Department of Statistics and Operations Research
Universitätsstraße 5/3/9
1010 Vienna, Austria
E-mail: andreas.futschik@univie.ac.at

ABSTRACT

An important problem in genetics is detecting single nucleotide polymorphisms (SNPs), sites along the genome at which a population shows variation. The focus is on the detection of rare variants. Pooling individuals allows us to increase the probability that a rare variant appears in the sample. However, as the pool size increases, the mean number of reads from an individual decreases, making it harder to distinguish reads of a rare variant from errors. This paper compares three statistical tests for detecting SNPs using data from pooled DNA samples.

Introduction

The genome consists of sequences made of 4 nucleotides (bases). At a majority of the sites in these sequences, each individual in a population has the same base. A site where there is variation is called a single nucleotide polymorphism (SNP). At such sites, in general, just two of the four bases appear. These variants are called alleles, the most common (rare) is termed the major allele (minor allele, respectively). We treat chromosomes, rather than members of a species, as individuals. However, our analysis can be generalized.

Since any reasonable test detects alleles of large frequency with power close to 1, we concentrate on the detection of low frequency alleles. Following Futschik and Schlötterer (2010), one may use the following test (referred to as the maximum test): accept that there is a minor allele if in any lane the number of reads for a non-major allele exceeds a given threshold. Ramsey and Futschik (2011) develop this work by specifying this threshold given the parameters of the sequencer and significance level required. An estimate of the power of this test is derived, which is used to find the asymptotically optimal pool size (optimal for detecting low frequency alleles).

One possible disadvantage of the test procedure described above lies in the fact that it only utilizes the maximum number of reads of a prospective minor allele from a lane and not the number of reads of a putative minor allele from each lane. For this reason, we compare the maximum test with two likelihood ratio tests, which use the number of reads of a putative minor allele from each lane. It is shown that for practical problems the maximum test is very effective. For more on the practical issues involved in gene pooling see Kenny *et al.* (2010).

Description of the Problem and a Simplified Model

Genome sequencers read DNA from a pool (of m individuals) placed in a lane. Suppose

we have k independent pools, i.e. the sample size is $n = km$. Consider a given site. If the same (large) amount of genetic material is taken from each individual, we may assume that the number of reads from an individual given there are r reads for that site in a lane has a binomial distribution with parameters r and $1/m$. Suppose that each read is incorrect with a small probability ϵ, independently of other reads. Also, suppose that only two alleles are possible, the major allele and the putative minor allele.

Let $\mathbf{R} = (R_1, R_2, \ldots, R_k)$, where R_i is the total number of reads for that site in lane i. It is assumed that the R_i are i.i.d. from the Poisson(λ) distribution. In addition, suppose good estimates of λ and ϵ are available for the gene sequencer used.

The major allele is inferred to be the one with the largest number of reads in the whole sample. As we are interested in detecting low frequency alleles, we may assume that for reasonable sample sizes the major allele is correctly identified with probability 1.

Denote the minor allele frequency at a given locus by p. We wish to define statistical tests of the following hypotheses.

H_0: The locus is not a SNP, i.e. $p = 0$.

H_A: $p = p_0$, where p_0 is some small positive value.

Let $\mathbf{X} = (X_1, X_2, \ldots, X_k)$, where X_i is the number of reads of the putative minor allele in lane i. Define A_i to be the number of individuals with the minor allele in lane i and $\mathbf{A} = (A_1, A_2, \ldots, A_k)$. Note that $A_i \sim \text{Bin}(m, p)$. We observe $\mathbf{X}$ and $\mathbf{R}$. The vector $\mathbf{A}$ can only be estimated from the data.

Consider the distribution of X_i conditional on the number of individuals with the minor allele and the total number of reads in lane i. Suppose that there is no-one with the minor allele in a pool. In this case, the number of reads of the prospective minor allele is simply the number of errors. Given $R_i = r$, X_i has a binomial distribution with parameters r and ϵ. It follows that X_i has a Poisson distribution with parameter $\lambda\epsilon$. Now suppose that there are a individuals with the minor allele in a pool of m individuals, $a \in \{1, 2, \ldots, m\}$. From our assumptions, the distribution of X_i given $R_i = r$ and $A_i = a$ is the binomial distribution with parameters r and $q(a)$, where

$$q(a) = \frac{a(1-\epsilon)}{m} + \frac{\epsilon(m-a)}{m}.$$

Hence, X_i has a Poisson distribution with parameter $\lambda q(a)$. Note that if ϵ is small in comparison to $1/m$, then for $a \geq 1$, $q(a) \approx a/m$.

Likelihood Ratio Tests

Based on the model presented in the previous section, we derive two likelihood ratio tests of the null hypothesis that there is no variation at a site. We treat the number of reads from a lane as an ancillary statistic (see Lehmann (1986), Chapter 10, Section 2). Hence, we compute our likelihoods conditional on $R_1, \ldots, R_k$. Under the alternative hypothesis, these likelihood functions also depend on the (unobserved) number of individuals with the minor allele in each lane. Firstly, we consider the likelihood of the number of reads of a prospective minor allele from a single lane. Under the null hypothesis that there is no variation at a site, we have

$$L_0(x_i | R_i = r_i) = \binom{r_i}{x_i} \epsilon^{x_i} (1-\epsilon)^{r_i - x_i}.$$

Similarly, under the alternative hypothesis

$$L_1(x_i | A_i = a_i, R_i = r_i) = \binom{r_i}{x_i} [q(a_i)]^{x_i} [1 - q(a_i)]^{r_i - x_i},$$

where $A_i \sim \text{Bin}(m, p)$. In the case $A_i = 0$, this is simply the likelihood under the null hypothesis.

A likelihood ratio test can be obtained by considering the information from all k lanes. This can be done in two ways: either we assume that the pools are obtained by sampling from a large population where the minor allele frequency is p, or we assume that the pools come from subpopulations with possibly different minor allele frequencies. In the second case, we simply infer the number of individuals with the minor allele in each pool. The denominator of the likelihood ratio is the same in both cases:

$$\prod_{i=1}^{k} L_0(x_i|R_i = r_i) = \epsilon^{\sum_{i=1}^k x_i}(1-\epsilon)^{\sum_{i=1}^k (r_i - x_i)} \prod_{i=1}^{k} \binom{r_i}{x_i}.$$

We start by considering the first strategy. In this case, the numerator is given by

$$\begin{aligned}\prod_{i=1}^{k} L_1(x_i|R_i = r_i) &= \prod_{i=1}^{k} \sum_{a_i=0}^{m} L_1(x_i|A_i = a_i, R_i = r_i) P(A_i = a_i) \\ &= \prod_{i=1}^{k} \sum_{a_i=0}^{m} \binom{m}{a_i} p^{a_i}(1-p)^{(m-a_i)} \binom{r_i}{x_i} [q(a_i)]^{x_i}[1-q(a_i)]^{r_i - x_i}.\end{aligned}$$

Setting $p = 0$, we obtain the likelihood under the null hypothesis. In the second case, where the a_i are considered as unknown parameters, we set

$$\prod_{i=1}^{k} L_1(x_i|A_i = a_i, R_i = r_i) = \prod_{i=1}^{k} \binom{r_i}{x_i} [q(a_i)]^{x_i}[1-q(a_i)]^{r_i - x_i}.$$

The likelihood under the null hypothesis is obtained by setting $a_1 = \cdots = a_n = 0$. We now maximize the numerator and compute the likelihood ratio. In the first case, this leads to

$$(1) \quad LR_1 = \frac{\max_p \prod_{i=1}^{k} \sum_{a_i=0}^{m} \binom{m}{a_i} p^{a_i}(1-p)^{(m-a_i)} [q(a_i)]^{x_i}[1-q(a_i)]^{r_i - x_i}}{\epsilon^{\sum_{i=1}^k x_i}(1-\epsilon)^{\sum_{i=1}^k (r_i - x_i)}}.$$

The second approach leads to the somewhat simpler expression

$$(2) \quad LR_2 = \frac{\prod_{i=1}^{k} \max_{a_i} [q(a_i)]^{x_i}[1-q(a_i)]^{r_i - x_i}}{\epsilon^{\sum_{i=1}^k x_i}(1-\epsilon)^{\sum_{i=1}^k (r_i - x_i)}},$$

where maximization can be carried out separately for each pool. The corresponding likelihood ratio test statistics are $2\log(LR_1)$ and $2\log(LR_2)$.

Likelihood ratio tests are very popular, and under suitable regularity conditions, their asymptotic properties are well understood. Indeed, in such a situation the null distribution of $2\log(LR_1)$ would be approximately chi-square with one degree of freedom and that of $2\log(LR_2)$ a χ^2_k distribution (see for instance Chapter 4 in Davison (2008)). However, the regularity assumptions on which this approximation is based are not satisfied here, as the null parameters are at the boundary of the parameter space and the a_i only vary on a discrete set of possibilities.

The Maximum Test

Consider the test statistic $U = \max_{1 \le i \le k} X_i$, i.e. U is the maximum number of reads of a putative minor allele in a lane. Under H_0, U is the maximum of independent observations from the Poisson($\lambda\epsilon$) distribution. The critical value for the test, u_k, is the smallest integer satisfying

$$P(U \le u_k|H_0) \ge 1-\alpha \Rightarrow P(X_i \le u_k|H_0)^k \ge 1-\alpha \Rightarrow P(X_i \le u_k|H_0) \ge \sqrt[k]{1-\alpha}.$$

Thus we can take the $\sqrt[k]{1-\alpha}$ quantile of the Poisson($\lambda\epsilon$) distribution as the critical value. We reject H_0 if and only if $U > u_k$.

Under H_A the number of minor alleles in the sample has a Bin(n, p_0) distribution. Ramsey and Futschik (2011) derive an approximation for the power of such a test for small p_0 and use this to define the asymptotically (as $p_0 \to 0$) optimal pool size. Note that the asymptotically optimal pool size is independent of p_0.

Results from Simulations

In order to estimate the power of the tests for given frequencies of the minor allele $p \in \{0.005,0.01,0.02,0.05\}$, error probabilities $\epsilon \in \{0.001,0.002,0.005,0.01\}$ and number of lines $k \in \{16,40,80,120\}$, we simulated ten thousand tests for each case with a maximum pool size of 10. It should be noted that the asymptotically optimal pool size in the problems considered is between three and six. The empirically determined optimal pool size for a particular test was taken to be the minimum pool size for which the maximum power was obtained.

It should be noted that the LR_1 test requires numerical maximization, in order to calculate the likelihood ratio statistic. The numerator of this expression (see Equation 1) was calculated at a grid of 20 equally spaced points $p = 0.003, 0.006, \ldots, 0.06$. Further simulations indicate that increasing the density of grid points did not not lead to any visible gain in power.

We carried out simulations under three models:

Model 1 Any errors in reading the major allele give the same allele (the minor allele, if one exists) and errors in reading the minor allele give the major allele.

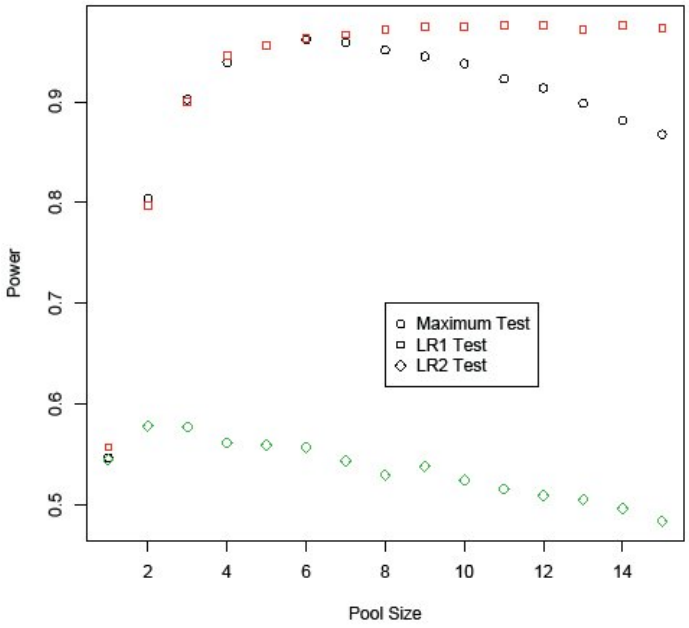
Model 2 Any errors in reading an allele always give the same allele, which is neither the major nor the minor allele.

Model 3 An error is equally likely to give any of the three other possible alleles.

Note that Model 1 corresponds to the statistical model described above. In Cases 2 and 3, more than two alleles can be observed at a site. As before, the major allele is assumed to be the allele with the largest number of reads in the whole sample. Using the maximum test, the putative minor allele is taken to be the non-major allele with the largest number of reads from a single lane. In the case of the likelihood ratio tests, we consider each of the 3 non-major alleles in turn. Each time, we classify the reads into two groups: (a) reads of the allele considered, (b) all other reads (we assume that these reads are of the major allele). We can then calculate the realization of the likelihood ratio statistic (LR_1 or LR_2) corresponding to each of the non-major alleles according to Equation (1) or (2), as appropriate. The putative minor allele is defined to be the one for which the maximum is achieved. If this statistic exceeds the appropriate critical value, then we accept that the putative minor allele is present. Note that it is possible to correctly reject H_0, but incorrectly infer which base is the minor allele. For such an error to occur, the test statistic corresponding to a non-present allele must exceed both the critical value for the test and the test statistic corresponding to the real minor allele. For this reason, we expect the probability of such an error to be smaller than the nominal significance level.

Figures 1 and 2 illustrate the results obtained for Model 1 using the appropriate asymptotic distributions to define the critical values for the likelihood ratio tests. Figure 1 is typical of cases for which several individuals with the minor allele are expected at the asymptotically optimal pool size. This version of the LR_2 test is clearly less powerful than the other two tests. The power of the LR_2 test was improved by using Monte Carlo simulation to estimate the critical value empirically. However, the power obtained using the other two tests remained superior. Hence, the LR_2 test was not applied to the other two models. For pool sizes below the asymptotically

Figure 1: ***Comparison of the power of the three tests for various pool sizes,*** $p = 0.01$***,*** $k = 80$***,*** $\epsilon = 0.001$***,*** $\alpha = 0.001$***. Based on 10,000 simulations.***



optimal pool size, the empirical power of the maximum and LR_1 tests were almost identical. For larger pool sizes, the power of the LR_1 test seems to increase slightly before plateauing, whilst the power of the maximum test slowly declines. Figure 2 is typical of the cases for which at most a couple of individuals with the minor allele are expected at the asymptotically optimal pool size. In this case, the maximum test works slightly better than the LR_1 test over the range of pool sizes considered. This is not particularly surprising, since in this case almost all the information regarding the presence of a minor allele is contained in the maximum number of reads of a putative minor allele from a lane.

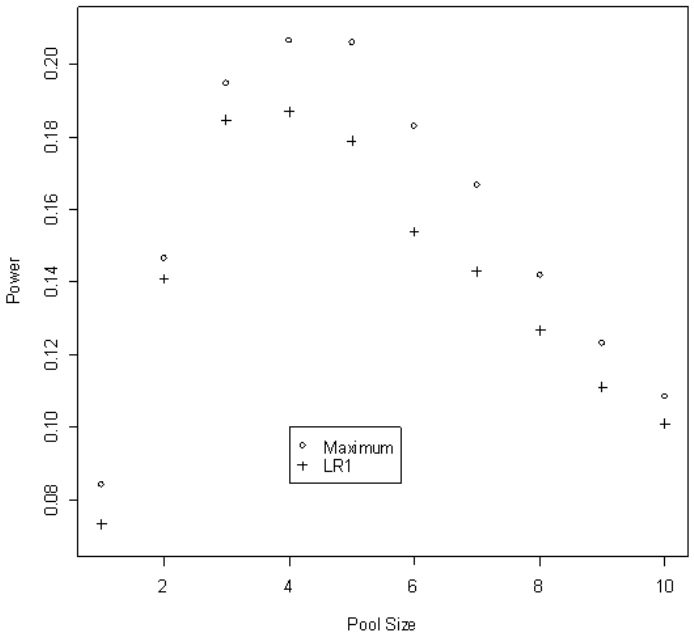
The results obtained under Model 2 are very similar to the results obtained under Model 1, except that the empirical powers of the maximum and LR_1 tests are slightly lower (mistakes in reading the major allele do not give the minor allele). The empirical probability of detecting the "wrong" minor allele is small compared the nominal significance level. Similar results are obtained under Model 3 using the maximum test. Under this model, the empirical power of the LR_1 test was clearly greater than that of the maximum test. However, further investigation indicated that the gain in power was comparable to the gain in power obtained using the maximum test when the error rate falls from ϵ to $\epsilon/3$ (the probability of a particular type of error). This suggests that it is relatively easy to adapt the maximum test when errors of various types can occur. In addition, in practice most misreads of an allele will give one particular allele.

In all cases, the tests are conservative (i.e. the empirical significance level is less than the nominal significance level). The empirically defined optimal pool sizes for the maximum test are robust to deviations from Model 1 and agree very well with the analytically derived pool sizes.

Conclusions

Compared to the maximum test, the LR_1 test can give a gain in power by using a larger pool size than the asymptotically optimal pool size when there are expected to be at least several

Figure 2: ***Comparison of the power of the maximum test and the LR_1 test for various pool sizes, $p = 0.005$, $k = 16$, $\epsilon = 0.01$, $\alpha = 0.001$. Based on 10,000 simulations.***



individuals with the minor allele in the sample. In practice, the pool size cannot be adapted to the expected number of individuals with a minor allele and using a relatively large pool size to improve detection of somewhat rare alleles will lead to a lower power in detecting rarer alleles. The results of the simulations indicate that the asymptotically optimal pool size is optimal or near optimal over a wide range of realistic parameters. These results suggest that using a fixed threshold to call SNPs is an effective and flexible procedure. One problem with this approach lies in the fact that λ varies from site to site. It is intended that future research will adapt the ideas presented in this paper to the analysis of real data.

Acknowledgements

D. M. Ramsey is grateful for the support of Science Foundation Ireland under the BIO-SI project (no. 07MI012)

REFERENCES

Davison, A. (2008) *Statistical Models*. Cambridge University Press, Cambridge.

Futschik A. and Schlötterer C. (2010) Massively parallel sequencing of pooled DNA samples - the next generation of molecular markers. *Genetics*, **186**: 207-218.

Kenny E. M., Cormican P., Gilks W. P., Gates A. S., O'Dushlaine C. T., Pinto C., Corvin A. P., Gill M., Morris D. W. (2010) Multiplex target enrichment using DNA indexing for ultra-high throughput SNP detection. *DNA Research* , doi: 10.1093/dnares/dsq029

Lehmann E. (1986) *Testing Statistical Hypothesis*, 2nd Ed. Springer, New York.

Ramsey D. M. and Futschik A. (2011) Optimal DNA pooling for the detection of single nucleotide polymorphisms. To appear in *Proceedings of the International Workshop on Statistical Modelling, Valencia, July 11-15, 2011.*